

Modernising and Harmonising Classification Standards

Submission to the Department of Infrastructure, Transport, Regional Development, Communications, Sport and the Arts

Associate Professor Andrew Moshirnia, Monash University

24 May 2026

I am an Associate Professor in the Department of Business Law and Taxation at Monash University, specialising in AI, technology and intellectual property law. I have published extensively on AI, video game design, and consumer protection, with particular focus on algorithmic manipulation, loot box monetisation mechanics, and the regulatory challenges posed by behavioural targeting in digital environments. I hold a Juris Doctor from Harvard Law School (2011) and a Doctor of Philosophy in Educational Technology from the University of Kansas (2008), where my doctoral dissertation examined the efficacy of information delivery systems in modified video games.

Overview

This submission engages with the *Update to Modernise and Harmonise Classification Guidelines 2025* (the Update), delivered by the Social Research Centre in partnership with Mettlesome, from the perspective of a researcher with sustained expertise in the design and consumer effects of video game monetisation systems. The Update proposes a welcome recalibration of the classification framework, removing the historically harsher treatment of video game interactivity relative to equivalent film content and replacing it with a replication principle under which only those game mechanics that replicate real-world harms attract heightened classification treatment.¹ The general direction of the Update reflects a sound and evidence-informed approach to harmonising Australia's classification framework.

This submission advances two specific concerns about gaps in the Update that risk undermining its protective purpose:

Concern 1: The Update correctly identifies, through Concept C04 (Interactivity classifications focus on replication), that gambling in a video game replicates the same psychological effect as real-world gambling. The revised guidelines should confirm that the otherwise relaxed approach² of C04 does not dilute the heightened classification treatment applicable to gambling-like variable-ratio loot box mechanics and instead should provide expressly for *uplift* to a mandatory minimum R18+ classification for games containing such mechanics. The Update's concept testing further supports this approach by confirming that advisory labelling is insufficient, finding that only one-third of 13-to-17-year-olds said a gambling warning would deter them from play.

Concern 2: The Update addresses artificial intelligence (AI) exclusively as a tool to assist classification processes (Concepts C13 and C14), while failing to address generative AI as a *content-generation risk* for the purposes of the classification guidelines themselves. This omission leaves a critical structural gap: as generative AI

¹ Social Research Centre and Mettlesome, *Update to Modernise and Harmonise Classification Guidelines 2025* (Department of Infrastructure, Transport, Regional Development, Communications, Sport and the Arts, 2025) 4.

² This would remove the current rule treating interactive video game content more harshly than equivalent film content, except where the mechanic directly replicates a real-world activity, such as gambling for reward, that carries equivalent harm regardless of medium. Ibid.

systems are deployed within games, films and other classified media to produce dynamic, personalised content at runtime, the existing classification framework, which is designed to assess static, pre-produced content, becomes structurally incapable of performing its protective function. Neither the Update's 'single-touch classification' principle (C10) nor its proposed mechanism for post-release updates (C11) can adequately address this risk. Accordingly, the revised classification guidelines should introduce a dedicated framework for generative AI content, requiring applicants to disclose the presence and parameters of runtime content generation at the point of classification, and establishing a classification-relevant definition of 'generative AI content' that encompasses dynamically produced material not present in any static, pre-classified version of the relevant game, film, or publication. This gap is rendered more acute by Concept C12's suggestion to make self-classification the industry default.

This submission elaborates on these concerns and proposes targeted and practical amendments to the Update.

INTRODUCTION

The Update represents a welcome and overdue effort to harmonise Australia's classification framework across media types and to ensure that the guidelines reflect contemporary community standards and harms research. The Update's framing is appropriate: it 'keeps what works, modernises what doesn't, and ensure[s] the system remains effective, inclusive, and trusted in a digital-first environment.' This submission engages with that stated purpose. Its concern is that two of the Update's most significant protective gaps are likely to cause material harm to the minors the classification framework is most directly designed to protect.

CONCERN 1: CONCEPT C04'S REPLICATION LOGIC REQUIRES CLASSIFICATION UPLIFT FOR VARIABLE-RATIO LOOT BOX MECHANICS

Australia's current classification framework draws a distinction between simulated gambling games, which carry a mandatory R18+ restriction, and games containing gambling-like paid loot boxes with variable-ratio reward schedules, which carry only an advisory M rating.³ This submission argues that distinction should not persist in light of the Update.

The Update's concept C04 proposes removing the historically harsher treatment of video game interactivity relative to equivalent film content but specifically carves out gambling mechanics from this relaxation. Gambling games should be assessed by reference to their real-world equivalents on the basis that 'gambling in a video game for a reward or incentive is, in effect, the same as gambling in the real-world.'⁴ Variable-ratio loot boxes deploy the same psychological mechanism, unpredictable reinforcement calibrated to maximise engagement, that underlies pokies and other regulated gambling products.⁵ The Update's replication logic therefore suggests a single, simplified classification outcome for both simulated gambling and gambling-like loot box games: a mandatory R18+ restriction. The gambling-harm research

³ Australian Classification Board, 'New Classifications for Gambling-like Content in Video Games' (*Australian Classification Website*, 2024) <<https://www.classification.gov.au/classification-ratings/new-classifications-for-gambling-content-video-games>>.

⁴ *Update* (n 1) 4.

⁵ Andrew Moshirnia, 'Precious and Worthless: A Comparative Perspective on Loot Boxes and Gambling' (2018) 20(1) *Minnesota Journal of Law, Science & Technology* 77, 80.

base reinforces this conclusion: research by the Australian Gambling Research Centre found strong associations between loot boxes and gambling harm across the surveyed population.⁶

The Update's own concept testing confirms that the advisory label is also ineffective with its intended audience. Only one-third of 13-to-17-year-olds, the group most at risk, said a gambling warning would make them less likely to play these games.⁷ If the advisory label does not deter the adolescent consumer from play, it cannot be said to perform any protective function for that consumer: the warning has been seen and disregarded, and the adolescent continues to engage with the harmful mechanic. The appropriate regulatory response to a protective measure that fails two-thirds of the minors it is designed to protect is not to retain it but to replace it with a legal restriction, the R18+ mandatory minimum, that already applies to related gambling content and does not depend on the consumer's voluntary response to information.

Moreover, the M advisory classification for gambling-like loot box mechanics is not working in part because game companies are simply not complying with the rating. Research conducted in September 2025 across 115 top-grossing mobile games found 20% non-compliance on Apple's App Store and 48% non-compliance on Google Play Store — with violations clustered in games carrying the lowest age ratings.⁸ On Google Play, 68% of G- and PG-rated top-grossing games were non-compliant. The maximum penalty of \$6,000 is negligible for games generating millions in revenue, creating perverse incentives to ignore rather than comply. An advisory classification that the industry demonstrably does not observe, and that carries no legal access restriction in any event, provides no meaningful protection. Most recently, a 2026 study documented Roblox's developer guidelines explicitly recommending loot box purchase prompts immediately after a player loses a match, conduct that exploits emotional states to engineer loot box purchases by children.⁹

The Update's replication logic and the compliance and deterrence evidence jointly compel uplift to a mandatory minimum R18+ classification for games containing variable-ratio loot box mechanics, consistent with the restriction that already applies to simulated gambling games and their real-world equivalents.

CONCERN 2: THE UPDATE DOES NOT ADDRESS GENERATIVE AI AS A CONTENT-GENERATION RISK

The Update's Treatment of AI Is Limited to Classification Process Tools

⁶ Nancy Greer, Cailiem Murray Boyle and Rebecca Jenkinson, *Harms Associated with Loot Boxes, Simulated Gambling and Other In-Game Purchases in Video Games: A Review of the Evidence* (Research Report, Australian Gambling Research Centre, Australian Institute of Family Studies, June 2022) <https://www.classification.gov.au/sites/default/files/documents/agrc_literature_review_final_20220906_accessible.pdf>.

⁷ *Update* (n 1) 17.

⁸ Marcus Carter et al, 'Investigating Mobile Games' Compliance with Australia's 2024 Mandatory Minimum Age Classifications Scheme for Gambling-Like Mechanics' (SSRN Working Paper, 2025) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5543999>.

⁹ Marcus Carter, Sanika Vekhande and Alisa-Jean Fifita, *Misleading and Deceptive Monetisation in Roblox: Evidence of False Pricing, Manipulative Design, and Consumer Harm in a Children's Digital Platform* (Sydney Games and Play Lab, University of Sydney, 2026) <<https://doi.org/10.25910/0gad-5g44>>; see also Chandni Gupta et al, *Playing the Player: Unfair Digital Gaming Practices and their Impact on Australians* (Consumer Policy Research Centre, March 2025) <<https://cprc.org.au/report/playing-the-player/>>.

The Update addresses artificial intelligence in two respects: Concept C13 proposes to update the classification database using ‘AI/natural language to aid in search functionality’, and Concept C14 proposes to expand ‘the use of automation and self-classification tools’ to support industry compliance.¹⁰ Both concepts address AI as an instrument for improving the *administration* of the classification system. Neither concept addresses the implications of AI for the *content* that classification is designed to assess.

The classification guidelines are designed to assess pre-produced, static content: a film with a fixed runtime, a game with a defined content set, a publication with a finite text. The classification process assumes that the content submitted for classification is the content the consumer will receive. That assumption is false for media that facilitates AI-generated content at runtime. The Update’s silence on this dimension of the AI question leaves a gap at the centre of the classification framework.

The Problem: Classification of Static Content Cannot Capture Runtime AI Generation

Generative AI systems deployed within games, interactive media, films and publications can produce text, imagery, dialogue, and narrative at runtime (that is, in response to user inputs) that was not present in any version of the content submitted for classification. This is not a hypothetical concern. Major gaming platforms already deploy large language model-powered non-player character (NPC) dialogue systems,¹¹ AI-generated procedural content systems that produce images and environments not authored by human developers,¹² and AI-powered chat features accessible to minors capable of generating content across a spectrum of themes and tones.¹³ Thus, the game that was classified by the Classification Board is, in material respects, not the game the consumer plays, because the consumer’s version of the game includes content that the Board never assessed.

Three categories of generative AI content risk are of immediate relevance to the classification framework and are not addressed in the Update. The first is runtime generation of content in a higher classification category than the classified product: a game classified at M may incorporate an AI dialogue system capable of generating R18+ material that the Classification Board never assessed because it was not present in any static version of the product submitted for classification. The second is AI-mediated generation of refused classification content:

¹⁰ *Update* (n 1) 5.

¹¹ Roberto Gallotta et al, ‘Large Language Models and Games: A Survey and Roadmap’ (2024) *IEEE Transactions on Games* <<https://doi.org/10.1109/TG.2024.3461510>>; Lucy O’Brien, ‘How Ubisoft’s New Generative AI Prototype Changes the Narrative for NPCs’ (Ubisoft News, 20 March 2024) <<https://news.ubisoft.com/en-us/article/5qXdxshJBXoanFZApdG3L/how-ubisofts-new-generative-ai-prototype-changes-the-narrative-for-npcs>>.

¹² Graham Todd et al, ‘Level Generation Through Large Language Models’ in *Proceedings of the 18th International Conference on the Foundations of Digital Games* (2023) <<https://doi.org/10.1145/3582437.3587211>>; Andrew Burnes, Seth Schneider and Jason Paul, ‘NVIDIA Redefines Game AI With ACE Autonomous Game Characters’ (GeForce News, 6 January 2025) <<https://www.nvidia.com/en-us/geforce/news/nvidia-ace-autonomous-ai-companions-pubg-naraka-blade-point/>>.

¹³ Alessandro Barbosa, ‘Square Enix Is Turning Slimes Into AI Chatbots In Dragon Quest X’, Gamespot (23 March 2026) <<https://www.gamespot.com/articles/square-enix-is-turning-slimes-into-ai-chatbots-in-dragon-quest-x/1100-6538960/>>; Aron Garst, ‘Darth Vader Could Be Just The Beginning Of Fortnite’s AI Implementation’, Gamespot (4 June 2025) <<https://www.gamespot.com/articles/darth-vader-could-be-just-the-beginning-of-fortnites-ai-implementation/1100-6532094/>>; eSafety Commissioner, AI chatbots and companions – risks to children and young people (18 February 2025) <<https://www.esafety.gov.au/newsroom/blogs/ai-chatbots-and-companions-risks-to-children-and-young-people>>.

generative AI systems may produce, on user prompt, content that would be refused classification if submitted as static content, and the RC category as currently framed applies to assessed content, not to a system's capacity to generate such content on request. The third is harm-calibrated personalisation: because generative AI systems can tailor content to a user's demonstrated preferences, emotional state, and engagement history, the content most harmful to the most vulnerable consumer, a minor exhibiting vulnerability markers, is precisely the content that will never appear in any static version of the product the Board classifies.

These concerns are well founded. Fortnite's AI-powered Darth Vader character, introduced in May 2025 and powered by Google's Gemini 2.0 Flash model, is a concrete example of a runtime generative AI system deployed within a game rated for children: within hours of launch, players had prompted the system to produce profanity and racial slurs that bypassed the content restrictions Epic Games had programmed into it.¹⁴ Several major chatbot or AI companion platforms have documented instances of users prompting AI-integrated features to generate content that bypassed intended content controls.¹⁵

The Update's Proposed Mechanisms Cannot Address This Gap

At first glance, two of the Update's own proposed mechanisms may appear capable of addressing runtime AI content generation, but neither is fully adequate.

Concept C11 (Permit updates to classifications) proposes that the database allow for classifications to be updated, retaining a record of the previous classification. This mechanism is reactive: it operates after harm has occurred and after a consumer has been exposed to content that the classification process failed to capture. For generative AI systems capable of producing harmful content on demand, post-hoc reclassification does not provide meaningful protection. More fundamentally, C11 is predicated on the existence of an identifiable, discrete content change (a patch, an update, a new version) that triggers the reclassification process. Generative AI systems do not produce such a discrete content change; the harmful content is generated in real time on user prompt, without any corresponding change to the product's static content set that would trigger C11's update mechanism.

Concept C10 (Single-touch classification) proposes that titles be classified once, only to be updated if content itself changes. This principle, which this submission generally supports for conventional static content, has the opposite effect for products incorporating generative AI systems: it locks in an original classification based on static content while the product's actual content set (that is, the content generated by the AI system at runtime) expands indefinitely in response to user interaction. A single-touch classification for a game with an AI dialogue system is, in practice, no classification of the AI-generated content at all.

¹⁴ Benji Edwards, 'The Empire Strikes Back with F-Bombs: AI Darth Vader Goes Rogue with Profanity, Slurs' (*Ars Technica*, 16 May 2025) <<https://arstechnica.com/information-technology/2025/05/the-empire-strikes-back-with-f-bombs-ai-darth-vader-goes-rogue-with-profanity-slurs/>>.

¹⁵ Michael Fire et al, 'Dark LLMs: The Growing Threat of Unaligned AI Models' (Working Paper, arXiv:2505.10066, 2025) <<https://doi.org/10.48550/arXiv.2505.10066>>; Senators Alex Padilla and Adam Schiff, Letter to the Federal Trade Commission regarding AI Chatbot Risks to Children and Teens (22 October 2025) <<https://www.padilla.senate.gov/newsroom/press-releases/padilla-schiff-urge-federal-trade-commission-to-address-ai-chatbots-risks-to-children-and-teens/>>; Senator Josh Hawley, Letter to Meta CEO Mark Zuckerberg regarding AI Chatbots Engaging Romantically with Kids (15 August 2025) <<https://www.hawley.senate.gov/wp-content/uploads/2025/08/Hawley-Letter-to-Meta-re-AI-Chatbots-Engaging-Romantically-with-Kids.pdf>>.

The gap is rendered more acute by Concept C12's proposal to make self-classification the industry default. Under C12, the Classification Board would be involved only 'in rare circumstances', with self-classification supported by 'evidence-based tools'. There are currently no self-classification tools capable of assessing the content generation capabilities of a deployed generative AI system. Expanding self-classification as a default at the same time as the guidelines fail to address generative AI content risks concentrating unassessed AI content risk in precisely the category of products, self-classified, that receives the least independent scrutiny.

Recommended Amendment

This submission recommends that the revised classification guidelines introduce a dedicated framework for generative AI-generated content. Specifically:

1. A classification-relevant definition of 'generative AI content' should be introduced across the guidelines for films, computer games and publications, in the following terms or equivalent:

Generative AI content: content, including text, imagery, dialogue, narrative, or audio, produced at runtime by an artificial intelligence system deployed within or accessible through the relevant film, computer game, or publication, that was not present in any static version of the product submitted for classification and that varies in response to user inputs, user behavioural data, or other runtime parameters.

2. Applicants for classification of products that incorporate or provide access to generative AI content systems should be required to disclose this at the point of classification application, including: (a) a description of the generative AI system deployed; (b) the content modalities the system is capable of producing; (c) the technical content filters or restrictions in place; and (d) whether the system's outputs are accessible to users under the age of 18. This requirement is consistent with the Update's broader commitment to transparency in classification decisions (Concept C18) and should be treated as a component of the classification application, not a post-hoc disclosure.
3. The Classification Board should be empowered to impose a classification condition on products containing generative AI systems, requiring applicants to notify the Board and re-submit for reclassification if the generative AI system is updated in a way that materially expands its content generation capabilities after the original classification decision. This provides a targeted complement to Concept C11's general post-release update mechanism, calibrated to the specific risk profile of AI content generation.
4. For generative AI systems accessible to users under 18, a mandatory minimum classification of M should apply unless the applicant can demonstrate that the system's content generation capabilities are technically restricted to content appropriate for the lower classification sought. This obligation should be assessed against the full range of content the system is theoretically capable of generating on user prompt, not merely the content it generates in standard gameplay or usage conditions. This minimum is in addition to, and does not displace, any higher mandatory minimum that applies to the product's static content under the existing classification guidelines.
5. Self-classification under Concept C12 should not be available for products incorporating generative AI content systems accessible to users under 18, pending the development of validated self-classification tools capable of assessing AI content generation capabilities. In the interim, such products should be required to seek

classification from the Classification Board, consistent with C12's existing reservation of Board classification for 'rare circumstances' that require independent assessment.

This framework is consistent with the approach developing in comparable regulatory contexts. The European Union's AI Act, which entered into force in August 2024 and began applying progressively from February 2025, establishes specific obligations for AI systems that generate synthetic content, including obligations to disclose the AI-generated nature of content and to implement content safeguards for systems deployed in contexts accessible to minors.¹⁶ The United Kingdom's Online Safety Act 2023 imposes obligations on platforms hosting content accessible to children, including content generated by AI systems, through safety duties that apply regardless of the mode of content production.¹⁷ The Update's own Phase 2 draft recommendations noted the opportunity to 'future-proof the system through ongoing monitoring, research on community standards, and exploration of artificial intelligence tools to support classification consistency.' This submission respectfully contends that future-proofing must extend beyond AI as a classification tool to AI as a classification subject.

CONCLUSION

The Update represents a constructive and evidence-informed effort to modernise Australia's classification framework. This submission supports its general direction. However, two gaps in the Update risk undermining its protective purpose – and these gaps lie in precisely the areas of greatest and most immediate risk to children and adolescents.

The Update's concept C04 correctly identifies that gambling in a video game replicates the same psychological effect as real-world gambling. The Update's own concept testing found that advisory gambling warnings are ineffective for the 13-to-17-year-old age group most at risk. The regulatory consequence that these two findings jointly compel is clear: legal restriction rather than advisory labelling for gambling-like variable-ratio loot box mechanics.

Generative AI systems are already deployed within games and other classified media to generate content at runtime that the Classification Board never assesses. The Update's silence on this dimension of AI, combined with its proposals to expand self-classification (C12) and adopt single-touch classification (C10), risks entrenching a structural classification gap that will grow as generative AI capabilities continue to develop. This submission recommends a targeted framework for generative AI content that can be implemented within the existing classification architecture without requiring new legislation.

Both recommendations are targeted and achievable. They are grounded in the Update's own evidence base and consistent with the direction of international regulatory developments. The classification framework's purpose, to help Australians make informed choices about what they and those in their care read, hear, watch and play, and to protect the community from unwanted exposure to material they may find harmful or offensive, cannot be served by an advisory label that two-thirds of adolescents say would not change their behaviour, nor by a classification process that cannot assess the content that consumers actually receive. This submission respectfully urges the department to address both gaps in the revised classification guidelines.

¹⁶ European Union, *Artificial Intelligence Act*, Regulation (EU) 2024/1689 of 13 June 2024 ('AI Act') art 5(1).

¹⁷ Online Safety Act 2023 (UK) <<https://www.legislation.gov.uk/ukpga/2023/50/contents>>.